Techno.COM, Vol. 20, No. 1, Februari 2021: 19-27

Pencocokan Berbasis Kata Kunci pada Penilaian Esai Pendek Otomatis Berbahasa Indonesia

Keyword Matching-Based on Short Essay Autograding in Indonesian

Nurul Chamidah[1], Mayanda Mega Santoni[2]
[1,2,]Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
E-mail: [1]nurul.chamidah@upnvj.ac.id, [2]megasantoni@upnvj.ac.id

Abstrak

Evaluasi dalam pengajaran dapat dilakukan melalui ujian. Ujian berupa esai dapat digunakan untuk mengevaluasi pemahaman sesuai konteks dan memiliki jawaban referensi. Sayangnya, jawaban dari esai ini membutuhkan waktu yang lebih banyak untuk dievaluasi dan dapat terjadi inkonsistensi dalam melakukan penilaian. Penelitian ini dilakukan untuk menganalisis performa untuk penilaian esai pendek otomatis berbahasa Indonesia untuk mengevaluasi jawaban yang berbentuk esai pendek. Sehingga, penilaian terhadap jawaban esai lebih konsisten dan dapat digunakan sebagai alternatif untuk penilaian dalam ujian online. Penilaian esai dilakukan dengan menghitung kecocokan antara jawaban uji dengan jawaban referensi, yakni dengan meilihat kata kunci dari masing-masing jawaban. Kata kunci diperoleh dengan melakukan praproses pada teks yakni dengan *case folding*, pembuangan *stopword*, *stemming*, dan tokenisasi. Setelah mendapatkan kata kunci untuk jawaban uji dan jawaban referensi, pada tahap *keyword matching* dilakukan pencocokan jawaban uji terhadap jawaban referensi. Hasil kecocokan antara jawaban uji dan referensi selanjutnya dihitung menjadi nilai pada tahapan *grading*. Nilai yang diperoleh dari *grading* selanjutnya dibandingkan dengan nilai uji sebagai evaluasi performa dengan menghitung *Mean Absolute Error* (MAE) dan *Pearson Correlation*. Hasil dari penelitian ini menunjukkan MAE untuk keseluruhan jawaban uji sebesar 0.25 dan korelasi antara nilai uji dengan nilai hasil *grading* sebesar 0.79.

Kata kunci: Penilaian Otomatis, Esai, Esai Pendek, Pencocokan Kata Kunci

Abstract

Teaching evaluation can be done through examinations. Essay exam models can be used to evaluate student understanding according to context and this type of exam has reference answers. Unfortunately, Essay's answer needs extra time to check and there can be inconsistencies in grading. This research was conducted to analyze keyword matching-based methods to evaluate short essays answers in Indonesian. Thus, essay answers grading is more consistent and can be used as an alternative in online exams grading. Essay grading is done by calculating keywords match between the test answer's keywords and the reference answers's keyword. Keywords are obtained by case folding, stopword removal, stemming, and tokenization. After getting the keywords for the test answers and the reference answers, we carried out a matching process between these keywords. The result of the match between the test answer and the reference is then calculated at the grading stage. The value obtained from grading is then compared with the test value as a performance evaluation by calculating the Mean Absolute Error (MAE) and Pearson Correlation. The results of this study indicate that the MAE for all test answers is 0.25 and the correlation between the test value and the grading result value is 0.79

Keywords: Autograding, Essay, Short Essay, Keywords Matching

1. PENDAHULUAN

Online learning atau *e-learning* saat ini banyak diterapkan dalam pembelajaran untuk menyampaikan materi pembelajaran, membagi materi yang berupa teks, gambar, dan video, mengumpulkan tugas, hingga digunakan untuk evaluasi seperti ujian dan kuis yang dapat diakses dari mana saja tanpa harus bertatap muka dengan pengajar.

Bentuk soal yang biasa digunakan untuk mengevaluasi suatu pembelajaran umumnya terdiri dari dua bentuk umum, yakni soal objektif dan soal esai. Soal objektif memiliki pilihan jawaban yang dapat berupa pilihan ganda, mencocokkan, dan benar salah. Sedangkan soal berbentuk esai tidak memiliki pilihan jawaban dimana penjawab harus menuliskan kalimatnya sendiri misal membuat karangan bebas yang tidak memakai kunci jawaban dan biasanya lebih memperhatikan tata bahasa, terdapat pula model soal esai yang meminta untuk menjelaskan atau mendeskripsikan sesuatu dimana penjelasan tersebut memiliki kunci jawaban.

Soal objektif lebih mudah diterapkan tapi lebih sulit untuk mengukur pengetahuan yang memerlukan jawaban berupa penjelasan, sedangkan soal esai dapat digunakan untuk mengukur kedalaman pengetahuan dari suatu pengajaran [1]. Meskipun esai dianggap lebih tepat untuk mengukur hasil kegiatan belajar, tapi bentuk soal ini memerlukan waktu yang lebih lama untuk mengevaluasi jawaban dari pada evaluasi pada soal yang berbentuk objektif. Masalah lain pada evaluasi soal ini adalah masalah konsistensi baik pada penilai yang sama maupun penilai yang berbeda [2]. Evaluasi soal esai oleh satu orang, bisa terjadi inkonsistensi bila dalam melakukan evaluasi dilaksanakan pada waktu yang terpisah. Apalagi bila evaluasi dilaksanakan oleh orang yang berbeda-beda bisa mendapatkan hasil nilai yang berbeda pada jawaban yang sama. Berdasarkan penjelasan tersebut, maka perlu dibuat suatu evaluator untuk menilai esai secara otomatis agar penilaian lebih konsisten.

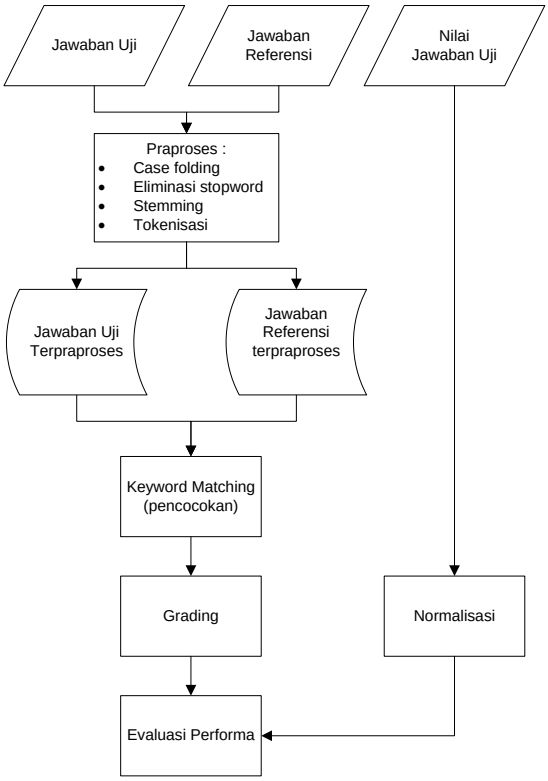
Sedangkan penilai esai pendek otomatis telah dilakukan oleh beberapa penelitian sebelumnya. Penilaian esai pendek menjadi 5 kelompok utama [3] yang terdiri dari klasterisasi, similaritas dokumen, pembelajaran mesin, ekstraksi informasi dan pencocokan pola, serta pengolahan bahasa alami atau *Natural Language Processing* (NLP). Penelitian-penelitian yang menilai esai dengan melihat struktur kalimat dan tata bahasa (NLP) dalam Bahasa Inggris seperti *Cambridge Learner Corpus-First Certificate in English Exam* (CLC-FCE) [4] mengevaluasi esai dalam Bahasa inggris yang melakukan pengecekan terhadap kesalahan linguistik termasuk deteksi kesalahan *grammar*. Penelitian [5] menggabungkan antara NLP dengan pembelajaran mesin dengan algoritma *Support Vector Machine* (SVM) dengan melakukan pelatihan untuk membangun model SVM.

Sedangkan penelitian-penelitian yang menggunakan Bahasa Indonesia telah banyak dilakukan sebelumnya. Namun, belum ada metode yang paling tepat untuk digunakan karena pada dasarnya dalam sistem penilai otomatis harus mampu memberikan penilaian jawaban sedekat mungkin dengan penilaian yang dilakukan oleh evaluator manusia dan sumber daya untuk Bahasa Indonesia tidak sebanyak sumber daya dalam Bahasa Inggris [6].

Penelitian penilai esai otomatis dalam Bahasa Indonesia antara lain menggunakan *Automated Essay Grading* menggunakan *Latent Semantic Analysis* (LSA) yang memetakan kata kunci dari jawaban uji dengan kata kunci dari jawaban referensi menjadi suatu matriks [2][7]. Penelitian lainnya menggunakan teknik *Vector Space Model* dengan membobotkan kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) dan menghitung *cosine similarity* antara jawaban uji dan jawaban referensi, yakni dengan mengambil kata kunci dari jawaban-jawaban tersebut [8], selain *cosine similarity*, penelitian [9] menggunakan *K-Nearest Neighbour* (KNN) untuk menghitung kedekatan antara kata kunci jawaban uji dan referensi. *Manhattan distance* dan *dice similarity* digunakan dalam penelitian [10] untuk menghitung kesesuaian jawaban uji dan referensi.

Pada kasus esai pendek, teknik dengan pembobotan kata dirasakan kurang efektif mengingat kata-kata yang muncul dalam esai pendek tidak banyak muncul secara berulang. Kemudian teknik dengan pembelajaran mesin memerlukan data latih dan data uji yang artinya membutuhkan cukup banyak data yang tersedia dalam suatu domain yang spesifik. Maka dari itu,

dalam penelitian ini kami akan mengembangkan suatu penilai esai pendek otomatis untuk mengevaluasi hasil ujian yang berupa esai pendek dengan memanfaatkan kata kunci yang diekstrak dari jawaban referensi dan jawaban uji. Penilaian sistem didasarkan pada jumlah kata kunci yang sesuai antara kedua jawaban tersebut.



Gambar 1. Metodologi Penelitian

2. METODE PENELITIAN

Gambar 1 menunjukkan metodologi penelitian yang digunakan dalam penelitian ini. Data yang diperlukan berupa jawaban uji, jawaban referensi (kunci jawaban), dan penilaian setiap jawaban oleh evaluator yang kami sebut sebagai nilai uji. Data teks jawaban uji dan jawaban referensi dipraproses dengan *case folding*, tokenisasi, eliminasi stopword, dan stemming. Setelah selesai dipraproses kata kunci dari jawaban uji dan jawaban referensi akan dicocokkan dan diberikan suatu nilai kecocokan yang merupakan nilai jawaban uji yang diberikan sistem. Proses terakhir dilakukan evaluasi yakni membandingkan nilai kecocokan dengan nilai jawaban uji untuk mengetahui seberapa baik kinerja sistem.

2.1. Data

Data yang diperlukan dalam penelitian ini terdapat 3 jenis, yakni jawaban yang akan diuji, jawaban referensi nilai dari jawaban uji dari evaluator, serta nilai maksimum dari soal. Data diperoleh dari ujian dengan bentuk esai pendek dan memiliki kunci jawaban pada mata kuliah pengantar basis data.

Data uji yang digunakan dalam penelitian ini berasal dari 4 soal dengan masing-masing satu kunci jawaban referensi. Masing-masing soal dijawab oleh 38 orang, sehingga total jawaban yang akan digunakan dalam penelitian ini adalah 152 jawaban dan 4 kunci jawaban. Tabel 1 merupakan contoh dataset yang terdiri dari soal, nilai soal, jawaban referensi, jawaban uji, dan nilai uji.

Tabel 1. Contoh Dataset

Soal	Nilai Soal	Jawaban Referensi	Jawaban Uji	Nilai Uji
Keberadaan sistem basis data di dalam Sistem Informasi manajemen adalah mutlak. Analisis mengapa basis data penting dalam pengembangan sistem informasi manajemen!	10	Karena basis data merupakan salah satu komponen dalam pengembangan sistem informasi. Basis data berperan dalam menyediakan data atau informasi bagi pemakai.	Basis data penting dalam pengembangan sistem informasi manajemen, karena basis data merupakan komponen penting dalam menyediakan data atau informasi yang akan digunakan oleh banyak user sesuai dengan tugas dan fungsi, dan setiap user pasti menginginkan ketersediaan data dan informasi secara akurat serta berkualitas. Maka dari itu, basis data harus dibuat dengan tepat sebagai acuan dalam pembuatan program atau aplikasi dan pengembangan sistem informasi manajemen sehingga lebih mudah dan cepat.	10

2.2. Praproses

Praproses teks jawaban uji dan jawaban referensi dilakukan untuk mendapatkan kata kunci. Praproses ini dilakukan dalam empat tahap yakni: *case folding*, eliminasi *stopword*, *stemming*, dan tokenisasi.

2.2.1. Case Folding

Case folding adalah tahapan proses mengubah semua huruf dalam teks dokumen menjadi huruf kecil, serta menghilangkan karakter selain a-z [11]. Tabel 2 menunjukkan hasil *case folding* dari contoh *dataset* pada Tabel 1, yakni pada jawaban uji dan jawaban referensi yang akan di proses. Perubahan pada *case folding* ini, huruf besar menjadi huruf kecil dan tanda baca seperti titik dan koma juga dihilangkan.

Tabel 2. *Case folding* pada jawaban referensi dan jawaban uji

Jawaban referensi	Jawaban Uji
karena basis data merupakan salah satu komponen dalam pengembangan sistem informasi. basis data berperan dalam menyediakan data atau informasi bagi pemakai.	basis data penting dalam pengembangan sistem informasi manajemen karena basis data merupakan komponen penting dalam menyediakan data atau informasi yang akan digunakan oleh banyak user sesuai dengan tugas dan fungsi dan setiap user pasti menginginkan ketersediaan data dan informasi secara akurat serta berkualitas. maka dari itu basis data harus dibuat dengan tepat sebagai acuan dalam pembuatan program atau aplikasi dan pengembangan sistem informasi manajemen sehingga lebih mudah dan cepat

2.2.2. Eliminasi Stopword

Eliminasi *stopword* atau penghapusan *stopword* bertujuan menyaring kata-kata dengan nilai diskriminasi yang sangat rendah. *Stopword* diproses pada sebuah kalimat jika mengandung kata-kata yang sering keluar dan dianggap tidak penting[12]. *Stopword* merupakan kata yang memiliki fungsi, tapi tidak memiliki arti. *Stopword* ini dieliminasi atau dihapus untuk mendapatkan kata-kata yang bermakna saja. *List* kata yang dihapus menggunakan *list stopword* Tala [13]. Tabel 3 menunjukkan hasil eliminasi *stopword* yang dilakukan pada Tabel 2. Dapat dilihat beberapa kata seperti karena, dalam, atau, bagi pada jawaban referensi di eliminasi. Begitu pula *stopword* pada jawaban uji.

Tabel 3. Hasil Eliminasi *Stopword*

Jawaban referensi	Jawaban Uji
basis data merupakan salah satu komponen pengembangan sistem informasi basis data berperan menyediakan data informasi pemakai	basis data penting pengembangan sistem informasi manajemen basis data merupakan komponen penting menyediakan data informasi akan digunakan banyak user sesuai tugas fungsi setiap user menginginkan ketersediaan data informasi akurat berkualitas dari basis data dibuat tepat acuan pembuatan program aplikasi pengembangan sistem informasi manajemen lebih mudah cepat

2.2.3. Stemming

Stemming merupakan proses untuk mendapatkan *root* atau *stem* atau kata dasar dari suatu kata dalam kalimat dengan cara memisahkan masing-masing kata dari kata dasar dan imbuhannya baik awalan (prefiks) maupun akhiran (sufiks) [14]. *Stemming* bertujuan untuk menghapus prefiks, sufiks, maupun afiks. Dalam penelitian dilakukan *stemming* menggunakan algoritma

Nazief-Adriani. Algoritma Sastrawi atau Nazief -Adriani dikembangkan pertama kali oleh Bobby Nazief dan Mirna Adriani, algoritma ini didasarkan pada aturan morfologi bahasa Indonesia yang luas dan dikumpulkan menjadi satu kelompok, serta dienkapsulasi pada imbuhan yang diperbolehkan dan yang tidak diperbolehkan [14].

Stemming diperlukan untuk mengambil kata kunci berupa kata dasar supaya tidak terpengaruh oleh bentuk aktif, bentuk pasif, pembendaan dan sebagainya karena imbuhan pada kata dihilangkan seperti pada kata merupakan menjadi rupa, pengembangan menjadi kembang, berperan menjadi peran, dll.

Tabel 4. Hasil *Stemming*

Jawaban referensi	Jawaban Uji
basis data rupa salah satu komponen kembang sistem informasi basis data peran sedia data informasi pakai	basis data penting kembang sistem informasi manajemen basis data rupa komponen penting sedia data informasi akan guna banyak user sesuai tugas fungsi tiap user ingin sedia data informasi akurat kualitas dari basis data buat tepat acu buat program aplikasi kembang sistem informasi manajemen lebih mudah cepat

2.2.4. Tokenisasi

Tokenisasi adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan memotong kalimat menjadi kata dengan menggunakan pemisah dapat berupa spasi, tab, dan *new line* [11]. Sedangkan menurut [15] tahap tokenisasi merupakan proses pemecahan sebuah teks menjadi bentuk kata atau disebut sebagai token. Jadi dapat disimpulkan bahwa tokenisasi merupakan pemecahan kalimat di dalam sebuah teks menjadi bentuk kata tunggal. Tabel 5 menunjukkan contoh hasil pemotongan dari kalimat menjadi kata. Kalimat jawaban yang awalnya berupa kalimat dalam paragraf menjadi kata-kata yang terpisah satu sama lain. Pada tahap ini telah diperlukan kata kunci yang berupa kata dasar dan siap untuk digunakan untuk proses *matching*.

Tabel 5. Hasil Tokenisasi Jawaban Uji dan Jawaban Referensi

Jawaban referensi	Jawaban Uji
basis, data, rupa, salah, satu, komponen, kembang, sistem, informasi, basis, data, peran, sedia, data, informasi, pakai	basis, data, penting, kembang, sistem, informasi, manajemen, basis, data, rupa, komponen, penting, sedia, data, informasi, akan, guna, banyak, user, sesuai, tugas, fungsi, tiap, user, ingin, sedia, data, informasi, akurat, kualitas, dari, basis, data, buat, tepat, acu, buat, program, aplikasi, kembang, sistem, informasi, manajemen, lebih, mudah, cepat,

2.3. Keyword Matching

Proses *keyword matching* dilakukan pada jawaban mahasiswa terhadap jawaban referensi dimana *keyword* atau kata kunci diperoleh dengan mengambil kata unik hasil tokenisasi. Contoh implementasi pencocokan ini dapat dilihat pada Tabel 6. Pada tabel tersebut dapat dilihat bahwa jawaban uji memiliki kecocokan terhadap jawaban referensi sebanyak delapan kata yang meliputi kata: basis, data, informasi, kembang, komponen, rupa, sedia, sistem.

Tabel 6. Pencocokan Kata Kunci

Jawaban referensi	Jawaban Uji
basis, data, informasi, kembang, komponen, pakai, peran, rupa, salah, satu, sedia, system	acu, akan, akurat, aplikasi, banyak, **basis**, buat, cepat, dari, **data**, fungsi, guna, **informasi**, ingin, itu, **kembang**, **komponen**, kualitas, lebih, manajemen, mudah, penting, program, **rupa**, **sedia**, sehingga, sesuai, **sistem**, tepat, tiap, tugas, user

2.4. Grading

Implementasi *grading* dilakukan dengan menghitung jumlah *keyword* jawaban uji yang sesuai dengan *keyword* jawaban referensi. Pada Tabel 6 dapat dihitung jumlah kata kunci pada jawaban referensi (N) adalah 12, dan kata kunci pada jawaban uji yang cocok atau beririsan dengan jawaban referensi (n) sebanyak 8, yang dapat dilihat pada huruf yang dicetak tebal. Sehingga, jawaban uji pada contoh tersebut memiliki nilai 8/12 = 0.67. Pada grading ini, penilaian akan bernilai 0 jika tidak terdapat kesamaan keyword, dan 1 jika keyword dari jawaban uji dan referensi sepenuhnya beririsan.

2.5. Normalisasi data nilai

Sebelum evaluasi dilakukan dengan membandingkan grading dari sistem dengan nilai uji yang sebenarnya, terlebih dahulu dilakukan normalisasi data pada nilai uji. Normalisasi ini dilakukan menyamakan skala penilaian antara nilai uji sebenarnya dengan hasil *grading* oleh sistem yakni diseragamkan kedalam nilai antara 0 hingga 1. Normalisasi nilai ini dilakukan dengan membagi nilai uji dengan nilai maksimum soal dengan normalisasi *min-max* [16]. Karena nilai minimum soal adalah 0, maka normalisasi *min-max* dapat dilihat pada rumus 1. Teknik normalisasi *min-max* ini dianggap paling tepat untuk digunakan dalam penelitian ini karena penilaian dari jawaban memiliki *range* yang pasti. Selanjutnya nilai hasil normalisasi ini akan digunakan untuk mengevaluasi sistem.

$$n_{baru} = \frac{n_{uji}}{n_{soal}} \quad (1)$$

Dengan n_{baru} adalah nilai normalisasi yang akan dicari, n_{uji} adalah nilai dari jawaban yang diberikan oleh evaluator, dan n_{soal} merupakan nilai penuh jika jawaban uji merupakan jawaban yang tepat. Sebagai contoh dalam Tabel 1, nilai uji 10, dan nilai soal 10, maka dalam hal ini, nilai baru yang dinormalisasi menjadi 1/1 = 1.

2.6. Evaluasi performa

Evaluasi dilakukan dengan menghitung *Mean Absolute Error* (MAE) antara nilai yang diberikan sistem dengan nilai yang diberikan evaluator. Formulasi MAE dapat dilihat pada rumus 2 sebagai berikut [17] :

$$MAE = \frac{\sum|n_{sistem} - n_{uji}|}{Jum} \quad (2)$$

Dengan n_{sistem} merupakan nilai sistem, n_{uji} berupa nilai uji yang berasal dari evaluator, dan Jum adalah jumlah jawaban yang dievaluasi.

Selain dengan MAE, evaluasi juga dilakukan dengan menghitung korelasi, yakni dengan *Pearson Correlation*. Korelasi ini digunakan untuk menghitung kesepakatan atau kesesuaian antara nilai yang diberikan oleh evaluator manusia dengan nilai yang dihasilkan oleh sistem. Nilai korelasi dihitung dengan rumus 3. Korelasi merupakan perbandingan antara kovarian dengan perkalian standar deviasi dari nilai uji dan nilai hasil grading oleh sistem.

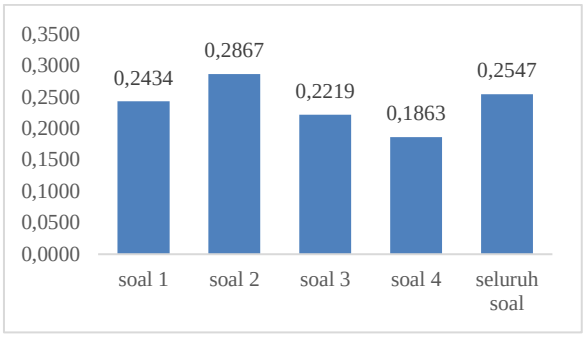
$$corr(n_{sistem}, n_{uji}) = \frac{cov(n_{sistem},\ n_{uji})}{stdev(n_{sistem}).stdev(n_{uji})} \quad (3)$$

Dengan $corr$ merupakan nilai korelasi antara nilai dari sistem dengan nilai uji, cov merupakan kovarian antara nilai sistem dan nilai uji, dan $stdev$ merupakan standar deviasi. Kriteria sukses dari sistem berdasarkan korelasi adalah, sangat baik jika korelasi > 0.75, baik jika korelasi antara 0.4 – 0,75, kurang jika korelasi < 0.4 [18].

3. HASIL DAN PEMBAHASAN

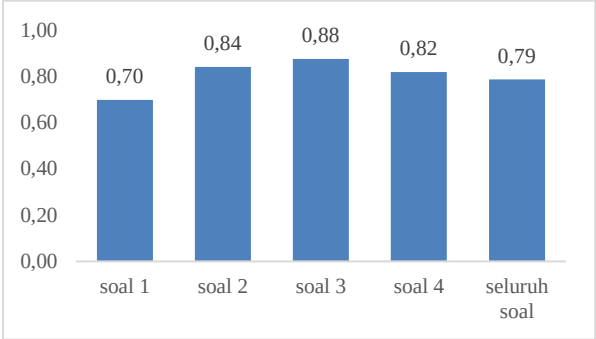
Percobaan dilakukan terhadap 4 soal dan 4 jawaban referensi dengan masing-masing soal terdapat 38 jawaban uji dengan nilai uji dan nilai soal diketahui. Pemrosesan yang dilakukan pada teks jawaban uji dan jawaban referensi meliputi praproses, *keyword matching*, dan *grading*. Evaluasi dilakukan dengan menghitung *Mean Absolute Error* (MAE) dengan membandingkan selisih nilai sistem dengan nilai uji dengan jumlah jawaban yang dievaluasi. Sebelum perhitungan MAE, normalisasi dilakukan pada nilai uji dengan mengubah *range* nilai menjadi antara nol hingga satu dengan rumus (1) sehingga nilai uji dan nilai hasil grading memiliki *range* yang sama.

Gambar 2 menunjukkan hasil perhitungan MAE pada penggunaan teknik *keyword matching* dalam penelitian ini. Dari gambar tersebut dapat dilihat bahwa *error* dari penilaian dengan sistem pada soal 1 sebesar 24.34%, soal 2 sebesar 28,67%, soal 3 sebesar 22.19%, dan soal 4 sebesar 18.63% atau untuk keseluruhan soal dengan menganggap jumlah uji adalah 35x4 = 140, yakni sebesar 25.47%.



Gambar 2. MAE

Korelasi antara *grading* sistem dengan nilai uji dapat dilihat pada Gambar 3. Dari gambar tersebut dapat dilihat bahwa soal 2, 3, dan 4 memiliki korelasi yang sangat baik karena memiliki nilai korelasi lebih dari 0.75, sedangkan soal 1 memiliki korelasi yang baik karena memiliki nilai antara 0.4-0.75. Sedangkan korelasi untuk seluruh jawaban adalah 0.79 yang masih sangat baik.



Gambar 3. Korelasi

Hal ini membuktikan bahwa secara umum metode yang digunakan dalam penelitian ini sudah memiliki korelasi yang baik dengan teknik penilaian yang dilakukan manusia meskipun perbedaan penilaian masih terjadi. Hal ini disebabkan oleh beberapa macam hal seperti penggunaan bahasa asing yang dapat dilihat pada Tabel 6 misal kata ‘user’ seharusnya berkorelasi dengan kata ‘pakai’ yang merupakan kata dasar dari ‘pemakai’. Selain itu, fleksibilitas dalam menggunakan istilah dengan makna yang mirip atau masih dapat diterima seperti dalam Tabel 1 jawaban referensi ‘Basis data berperan dalam menyediakan data atau informasi bagi pemakai’ sedangkan di jawaban uji ‘basis data merupakan komponen penting dalam menyediakan data atau informasi yang akan digunakan oleh banyak user sesuai dengan tugas dan fungsi’ dalam jawaban referensi terdapat kata ‘berperan’ sedangkan pada jawaban uji disebut sebagai ‘komponen penting’.

4. KESIMPULAN DAN SARAN

Penilaian esai otomatis tanpa pelatihan dapat dilakukan dengan menggunakan kata kunci, yakni dengan mencocokkan jawaban uji dengan jawaban referensi. Dari hasil eksperimen yang dilakukan diperoleh bahwa teknik ini mampu melakukan penilaian otomatis dengan hasil penilaian yang memiliki korelasi sangat baik dengan penilaian yang dilakukan oleh evaluator manusia. Meskipun demikian, *Mean Absolute Error* (MAE) dari sistem masih cukup besar yakni 0.2547 atau 25.47% yang menunjukkan penilaian oleh sistem masih memiliki perbedaan cukup signifikan dengan penilaian evaluator.

Beberapa hal yang mungkin dilakukan untuk pengembangan teknik penilaian esai pendek otomatis ini adalah dengan menggunakan mesin penerjemah untuk istilah asing atau dengan menyediakan jawaban alternatif, serta memanfaatkan tesaurus untuk mengatasi variasi kata yang digunakan sehingga dapat memperkaya kosa kata.

UCAPAN TERIMA KASIH

Terima kasih kepada Universitas Pembangunan Nasional Veteran Jakarta yang telah mendanai penelitian ini dalam skema Riset Dosen Pemula pada tahun anggaran 2020.

DAFTAR PUSTAKA

[1] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 2017, pp. 697–702.

[2] R. Adhitia and A. Purwarianti, "PENILAIAN ESAI JAWABAN BAHASA INDONESIA MENGGUNAKAN METODE SVM - LSA DENGAN FITUR GENERIK," *J. Sist. Inf.*, vol. 5, no. 1, p. 33, Jul. 2012.

[3] S. Roy, Y. Narahari, Y. Narahari, and O. D. Deshmukh, "A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers Melding Game Theory with Machine Learning View project Fair Allocation of Indivisible Goods View project A Perspective on Computer Assisted Assessment Techniques for Short Free-text Answers," *Springer*, vol. 571, pp. 96–109, 2015.

[4] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, vol. 1, pp. 180–189.

[5] A. Suresh and M. Jha, "Automated Essay Grading using Natural Language Processing and Support Vector Machine," *IJCAT-International J. Comput. Technol.*, vol. 5, no. 2, pp. 5–8, 2018.

[6] F. Rahutomo *et al.*, "Open Problems in Indonesian Automatic Essay Scoring System," *Int. J. Eng. Technol.*, vol. 7, no. 4.44, p. 156, Dec. 2018.

[7] R. B. Aji, Z. A. Baisal, and Y. Firdaus, "Automatic Essay Grading System Menggunakan Metode Latent Semantic Analysis E-78 E-79," *Semin. Nas. Apl. Teknol. Inf.*, vol. 2011, no. Snati, pp. 1–9, 2011.

[8] J. Zeniarja, A. Salam, and I. Achsanu, "Sistem Koreksi Jawaban Esai Otomatis (E-Valuation) dengan Vector Space Model pada Computer Based Test (CBT)," *Seri Pros. Semin. Nas. Din. Inform.*, vol. 4, no. 1, Apr. 2020.

[9] M. Jamaluddin, N. Yuniarti, A. Rahmani, and J. Hutahaean, "Aplikasi Penilaian Otomatis Ujian Esai Berbahasa Indonesia Menggunakan Algoritma K-Nearest Neighbor (Studi kasus MAN Cimahi)," *Pros. Ind. Res. Work. Natl. Semin.*, vol. 10, no. August 2019, pp. 314–324, Aug. 2020.

[10] F. Rahutomo, Y. P. Putra, and M. H. Ali, "Implementasi Manhattan Distance dan Dice Similarity pada Ujian Esai Daring Berbahasa Indonesia," *Semin. Inform. Apl. Polinema*, pp. 171–174, 2019.

[11] F. Pratama, "Rancang Bangun Aplikasi Peringkas Tkes Otomatis Artikel Berbahasa Indonesia Mengunakan Metode Term Frequency Inverse Document Frequency (TF-IDF)

dan K-mean Clustering," *Fak. Sains dan Teknol.*, Apr. 2014.

[12] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[13] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[14] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, 2017.

[15] A. Sukma, B. Zaman, and E. Purwanti, "Information Retrieval Document Classified with K-Nearest Neighbor," *Rec. Libr. J.*, vol. 1, no. 2, p. 129, Jan. 2018.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[17] G. Brassington, "Mean absolute error and root mean square error: which is the better metric for assessing model performance?," *Geophys. Res. Abstr.*, vol. 19, pp. 2017–3574, 2017.

[18] T. F. de C. Marshall and J. L. Fleiss, "Statistical Methods for Rates and Proportions.," *Stat.*, vol. 25, no. 1, p. 70, 1976.